Соловьев С.Ю.

Постановки задач современной информатики

park.glossary.ru/modern/

Задачи статистического анализа

2015 –2022

Статистический анализ – теория конструирования и применения методов выявления причинно-следственных связей.

Математическая статистика – раздел математики, в котором разрабатываются методы регистрации, описания и анализа данных наблюдений с целью получения вероятностно-статистических моделей случайных явлений.

Напоминание: **З а д а ч а**

Дано	Известно
Исходные данные	***Свойства исх. данных***

Алгоритм / Метод / Способ / Схема

Требуется	такое, что
Результирующие данные	***Свойства рез. данных***

VS.

Математическая статистика

Частоты

Выборки

Оценки характеристик

Теория вероятностей

Случайные величины

Характеристики СВ:

математическое ожидание,
дисперсия, **м**ода и др.

функции распределения

Задачи математической статистики

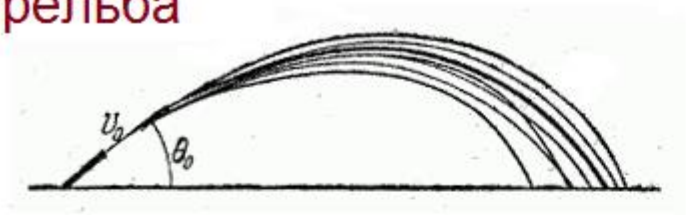
Дано	Известно
Выборка …	***Теория вероятностей***

Алгоритм / Метод / Способ / Схема

Требуется	такое, что
Оценки характеристик	***Теория вероятностей***

Ситуации с неопределенным исходом

Стрельба



Подбрасывание

Стохастические ситуации

Теория вероятностей

Поиск цивилизаций

Основные понятия

Стохастическая ситуация:
- непредсказуемость,
- воспроизводимость,
- устойчивость частот.

Вероятность событий ⟵ *lim*

vs. Аксиоматическое определение вероятности:

$P : \{ \text{События} \} \rightarrow [0,1]$ + A1 +A2 +A3 +A4

Событие = { Элементарные исходы }

Случайная величина ::=

функция : { Элемент.исходы } → { **x** | x - число }

Основные *понятия 2*

Случайные величины

Дискретные

Распределение вероятностей

$\{ (x_k, p_k) \mid k=1,\dots \}$

: $p_1+\dots+p_n = 1$

Непрерывные

Плотность вероятностей $f(x) \geq 0$

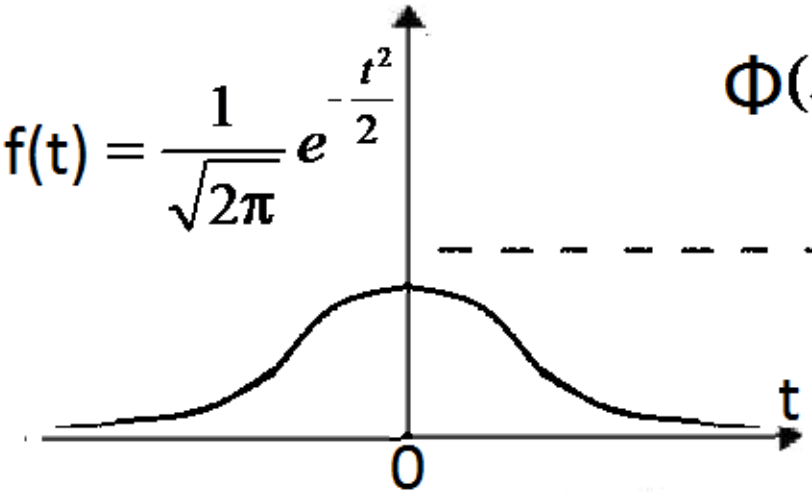
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Функция распределения

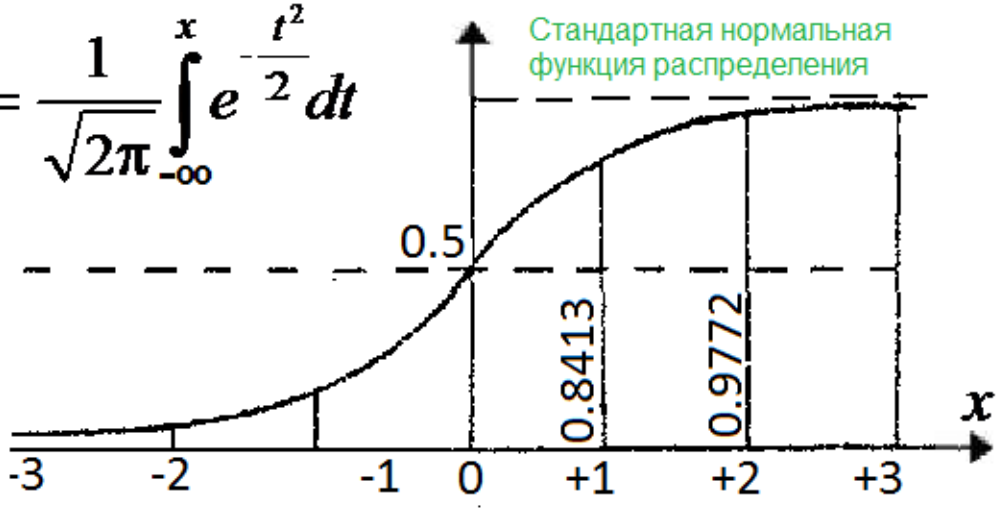
$$\mathbf{F(x)} = P(X < x) = \begin{cases} \sum_{k:x_k < x} P(X = x_k) & \text{если } X - \text{дискретная} \\ \int_{-\infty}^{x} f(t)dt & \text{если } X - \text{непрерывная} \end{cases}$$

Подзадачи ТВ и МС

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$



$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

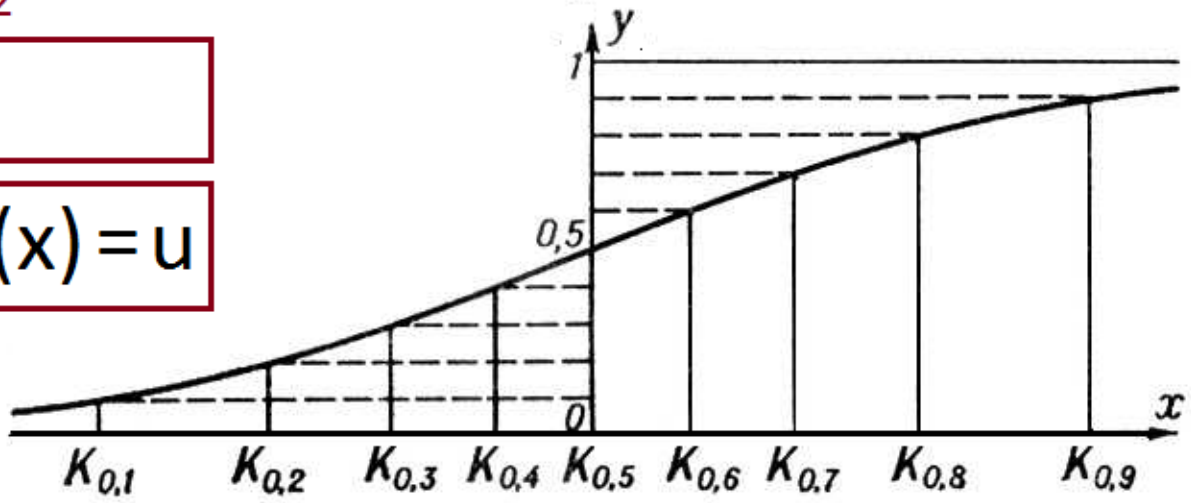


Подзадача 1

f, x	
Φ(x)	

Подзадача 2

f, u	
x	Φ(x) = u



Числовые характеристики СВ X

характеристики “центра”

Математическое ожидание

$EX = x_1p_1 + x_2p_2 + \dots x_np_n$ (диск.)

$EX = \int x\, f(x)\, dx$ // от $-\infty$ до $+\infty$ (непр.)

не каждая СВ имеет матем.ожид.

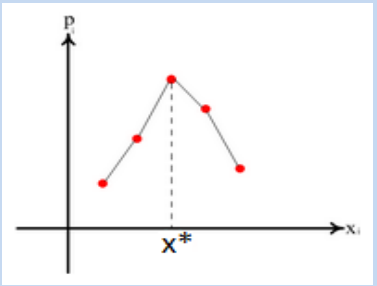
Медиана med X ::=

$x^* \in R : P(X \le x^*) \ge 0.5$ & $P(X \ge x^*) \ge 0.5$

существует, но не единственная

Мода дискретной СВ mod X ::=

$x^* \in$ знач. $X : P(X = x_k) \le P(X = x^*)$



Числовые характеристики СВ X

характеристики “разброса”

Дисперсия

$DX = E(X-EX)^2$

Среднеквадратичное отклонение

$\sigma = \mathrm{sqrt}(DX)$ // $\sigma^2 = DX$

Другие

$E|X - \mathrm{med}\, X|$

интерквартильный размах

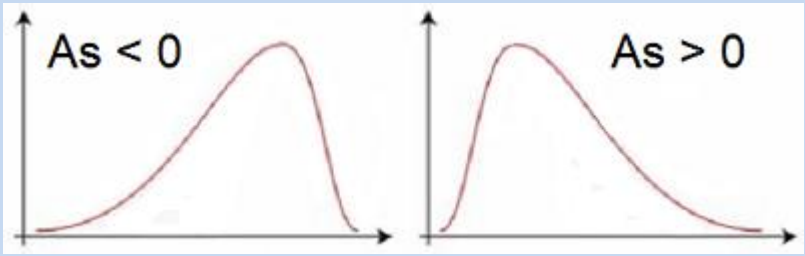
Числовые характеристики СВ X

характеристики “формы”

Коэффициент асимметрии

$$E\left(\frac{X-EX}{\sqrt{DX}}\right)^3$$

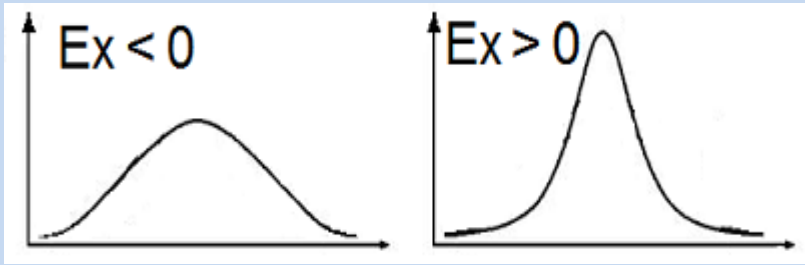
если существует $E|X|^3$



Коэффициент эксцесса

$$E\left(\frac{X-EX}{\sqrt{DX}}\right)^4$$

если существует $E|X|^4$

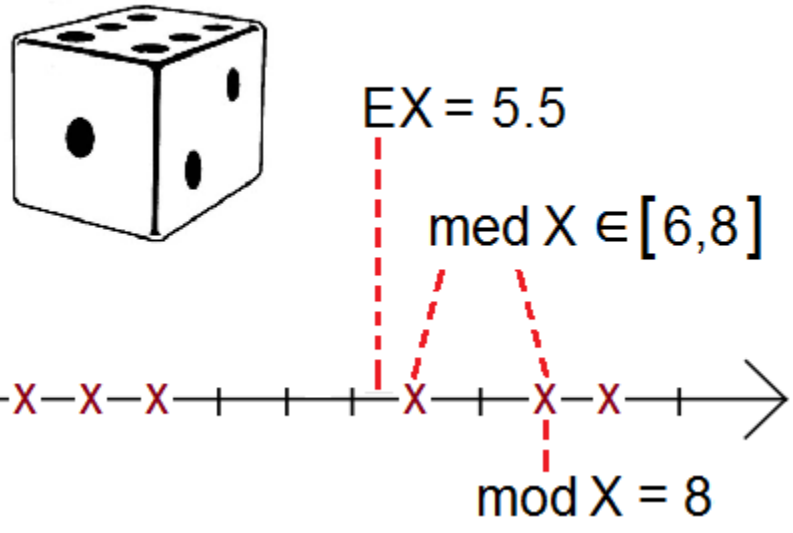


Числовые характеристики СВ X

пример

N.B. вероятности

X :	0	1	2	6	8	9
	0.1	0.1	0.15	0.15	0.3	0.2



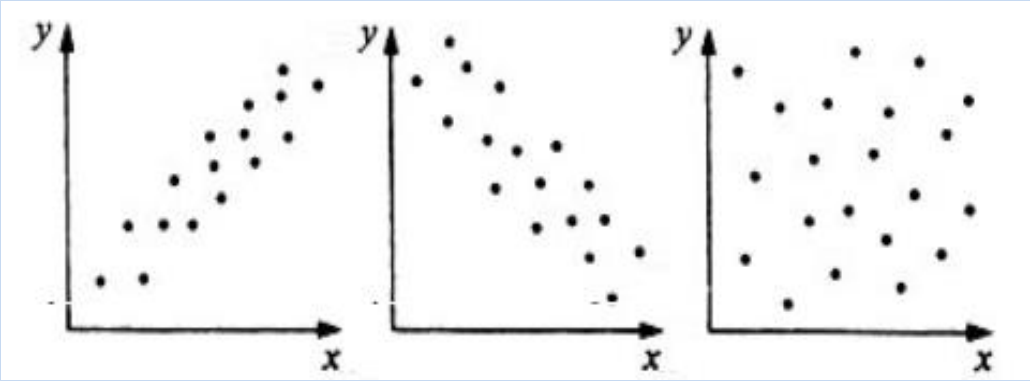
$DX = 11.25 \quad \sigma = 3.35$

Числовые характеристики взаимозависимости СВ X И Y

Ковариация $\mathrm{cov}(X,Y) = E(X-EX)(Y-EY)$

Коэффициент корреляции [-1,+1]

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{DX\,DY}}$$



$\rho > 0$ $\rho < 0$ $\rho \approx 0$

Основные понятия математической статистики

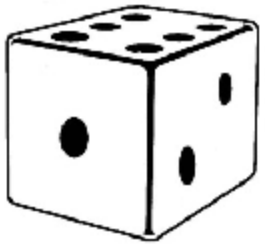
Выборка = $X_1, \ldots, X_n$, где X_i – (а) независимые и (б) одинаково распределенные СВ.

n – объем выборки

Выборка объемом 14

0, 8, 2, 8, 9, 1, 9, 0, 9, 9, 8, 1, 2, 0

Генеральная совокупность



X :	0	0.1
	1	0.1
	2	0.15
	6	0.15
	8	0.3
	9	0.2

Выборочные числовые характеристики

Выборочное среднее $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\approx EX$

Выборочная дисперсия $S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$ $\approx DX$

Выборочное среднеквадратичное отклонение

$S = \sqrt{S^2}$ $\approx \sigma$

Выборочный центральный момент порядка r

$m_r = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^r$

Закон больших чисел & Центральная предельная теорема

$X_1, \ldots, X_n$

-- независимые одинаково распределенные СВ

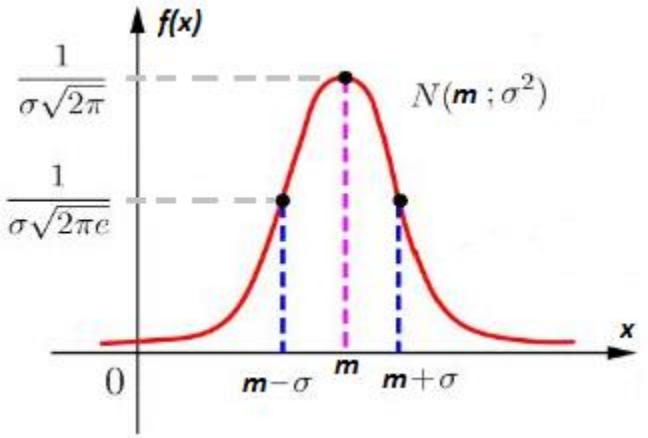
$$P\left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - EX \right| < \varepsilon \right) \geq 1 - \frac{DX}{n\varepsilon^2}$$

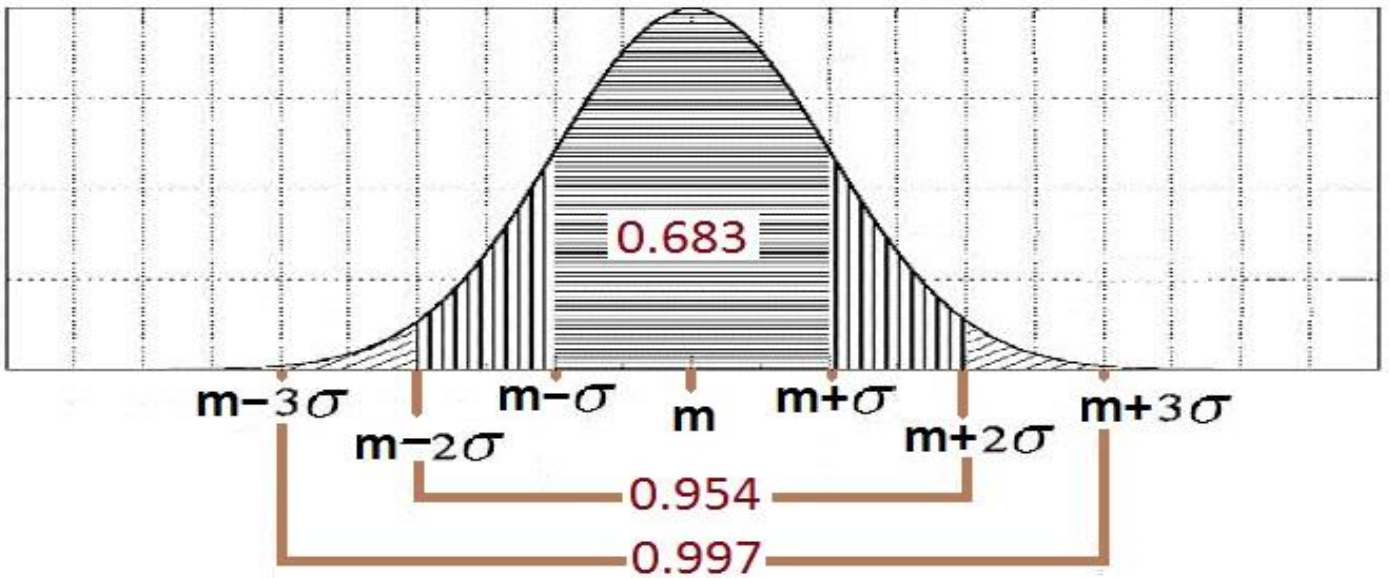
$$P\left(\sum_{k=1}^{n} X_k < x \right) \approx \Phi\left(\frac{x - nEX}{\sigma\sqrt{n}} \right)$$

Нормальное распределение n=1

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

m = EX



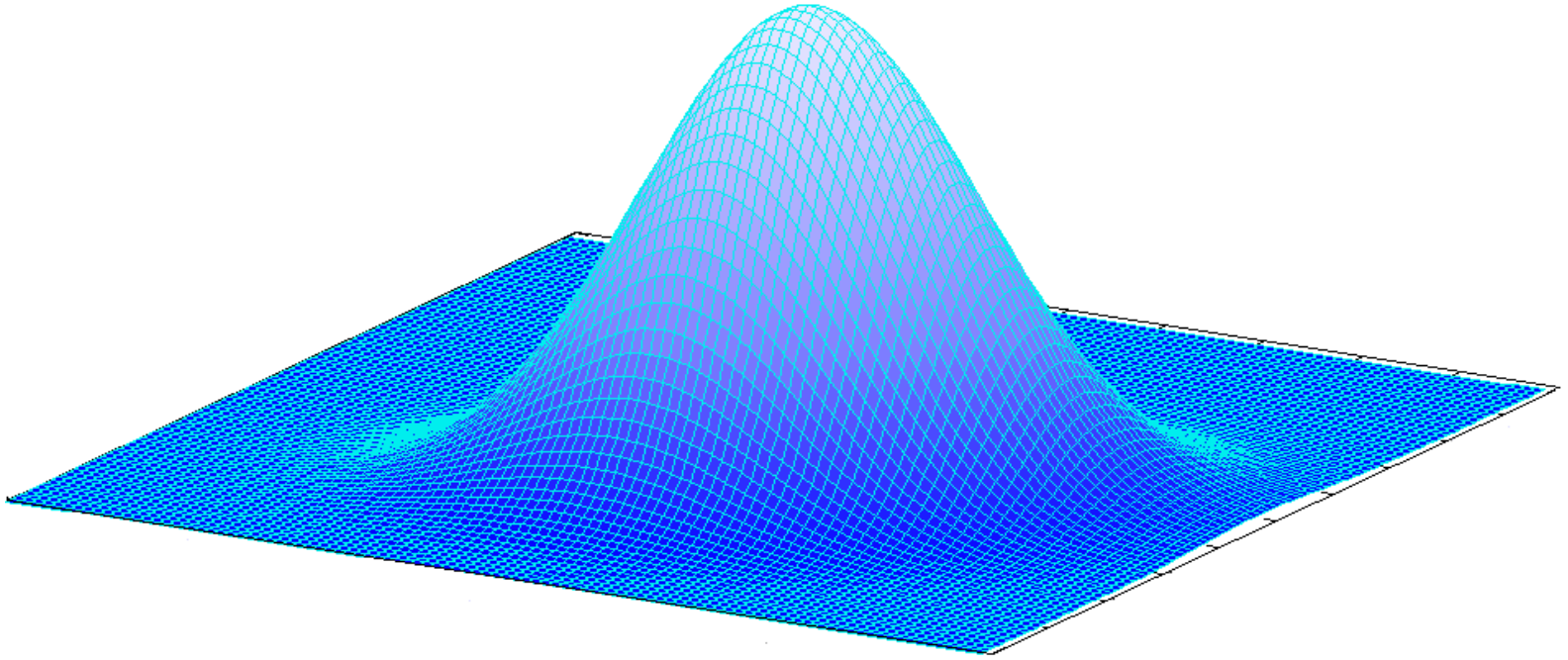


Нормальное распределение n > 1

Инфо: *многомерная плотность нормального распределения*

$$N(x;\mu,\Sigma) = (2^n\pi^n|\Sigma|)^{-0.5}\exp(-0.5(x-\mu)^T\Sigma^{-1}(x-\mu))$$

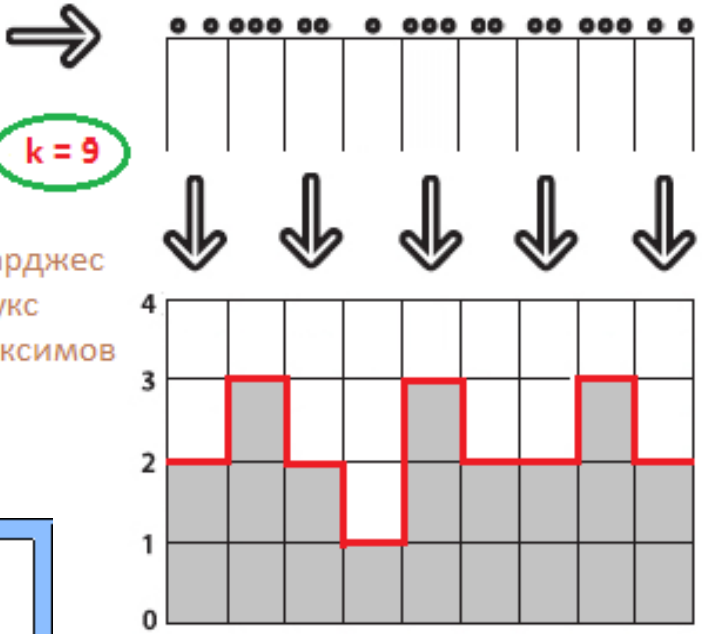
$\mu \in R^n$ – *матем.ожидание*, $\Sigma \in R^{n\times n}$ – *ковар.матрица*



Выборка >> Гистограмма

$X_1, \ldots, X_{20}$

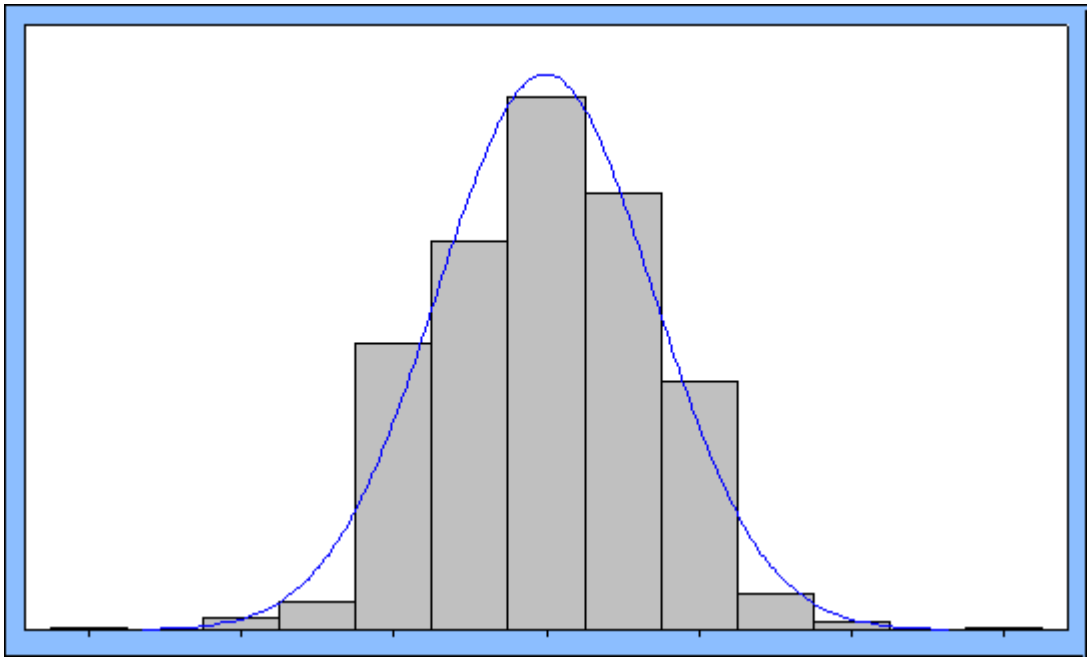
n = 20



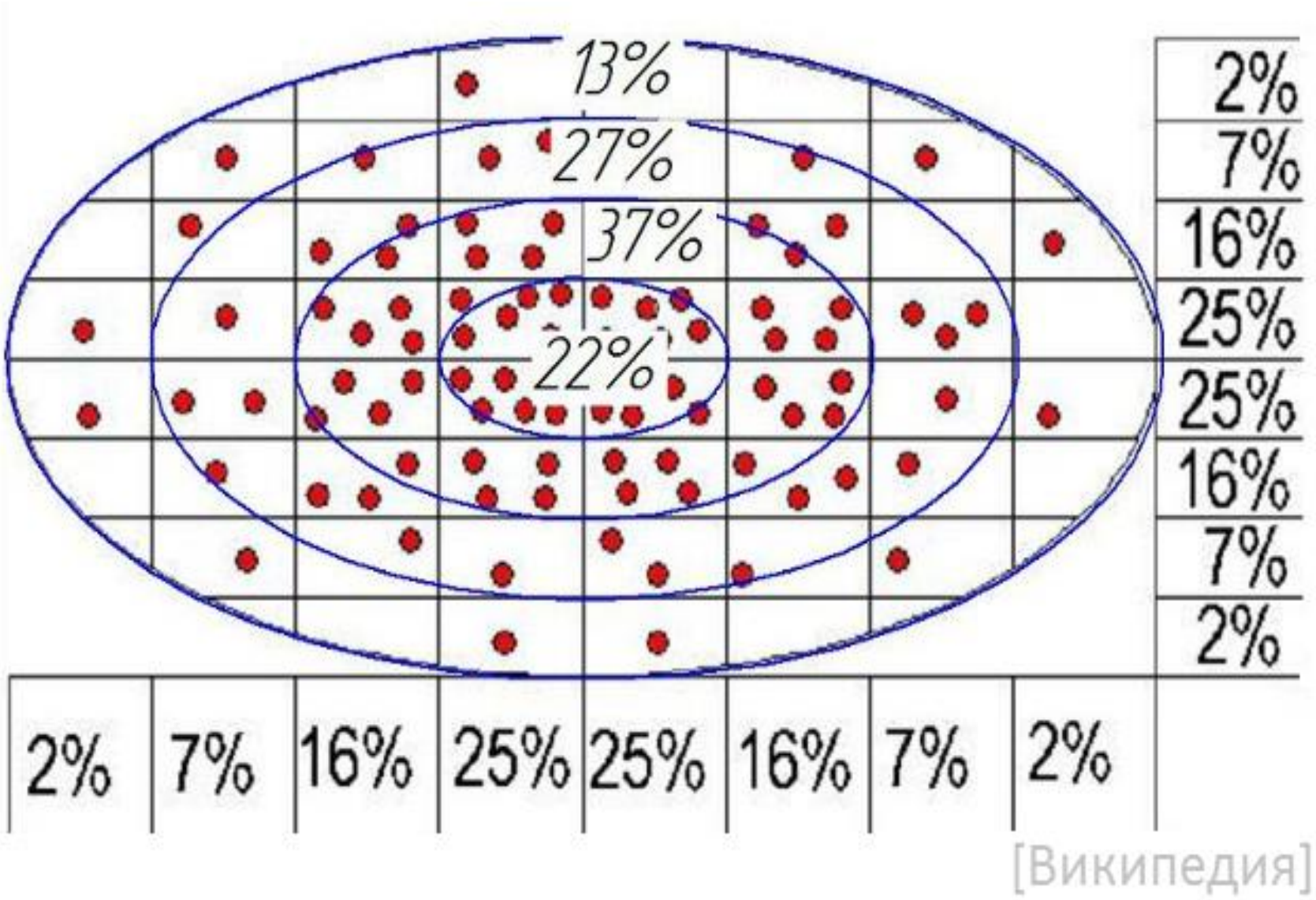
k = [3.3 lg n +1] – Старджес

k = [5 lg n] – Брукс

k = [1.72 $n^{1/3}$] – Максимов

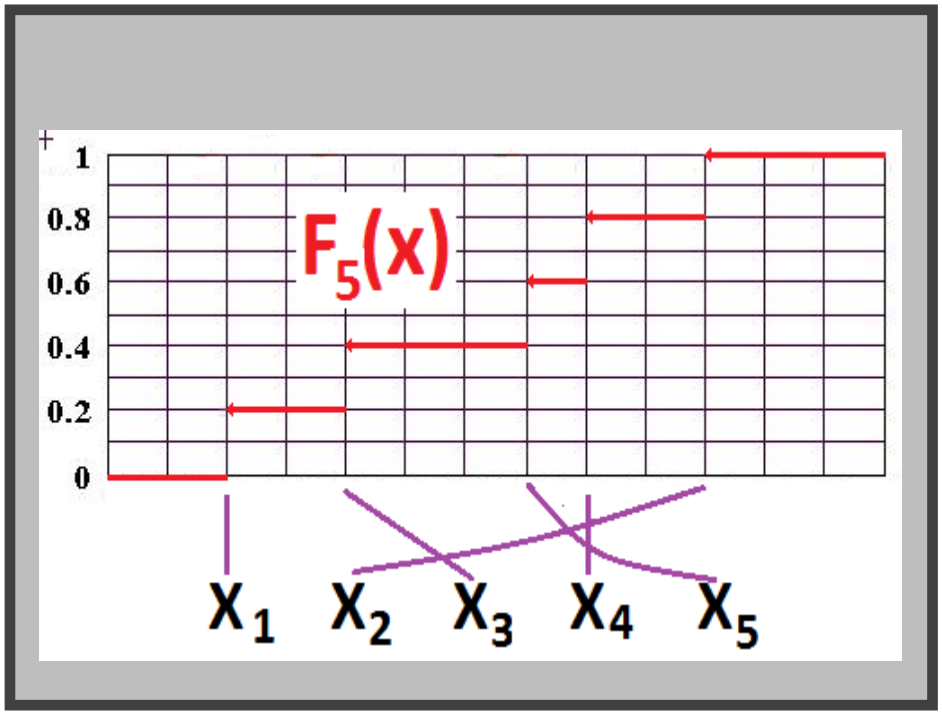


Выборка >> Эллипс рассеяния

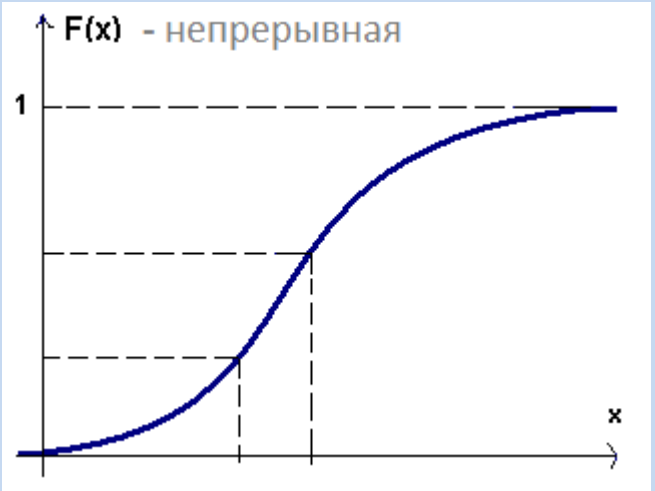


[Википедия]

Репрезентативность выборки



Генеральная совокупность



$$P(\max_x |F_n(x) - F(x)| \le \varepsilon) \ge \gamma$$

$$\Rightarrow K(\gamma) \approx \varepsilon \times sqrt(n)$$

$\gamma = 0.95 \quad \varepsilon = 0.035 \quad n \approx 1500$

$\gamma = 0.95 \quad \varepsilon = 0.001 \quad n \approx 1849600$

Основные классы задач математической статистики

1. Определение вероятности события.

2. Оценивание числовых характеристик СВ:

– точечное оценивание;

– интервальное оценивание;

3. Нахождение закона распределения СВ.

4. Проверка статистических гипотез.

5. Определение регрессионных зависимостей.

Точечные статистические оценки

Точечная статистическая оценка ::=

Функция : Выборка → Число,

где **Число θ –** оценка χарактеристики СВ.

Свойства оценок:

-- состоятельность: $\theta \to \chi$ *при* $n \to \infty$ *по вер.*

-- несмещаемость: $E\theta = \chi$

-- робастность: *устойчивость к выбросам*

-- эффективность: *min дисперсии*

Точечные статистические оценки

Выборочное среднее –

(а) состоятельная, (б) несмещенная и (в) неробастная

оценка математического ожидания СВ.

Выборочная дисперсия –

(а) состоятельная, (б) смещенная и (в) неробастная

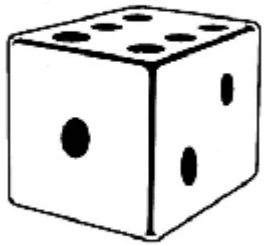
оценка дисперсии СВ.

$$\Rightarrow \quad \text{несм.}\ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Точечные статистические оценки

пример

Объем	500	5000	32000
Частоты	0.11	0.11	0.10
	0.09	0.09	0.10
	0.16	0.16	0.15
	0.12	0.14	0.15
	0.34	0.30	0.30
	0.18	0.20	0.20
Выб.средн	5.46	5.44	5.46
Выб.дисп.	11.54	11.44	11.35



X :	0	0.1
	1	0.1
	2	0.15
	6	0.15
	8	0.3
	9	0.2

EX = 5.5

DX = 11.25

Задача нахождения доверительного интервала

для вероятности события при большом объеме выборки

Дано числа n, p*, γ	n – объем выборки, $1/(n \times \mathrm{sqrt}(n)) \approx 0$; p – вероятность события A; (обозн.) p* – отн.частота события A; γ – доверительная вероятность;

Алгоритм*)

числа θ_1 и θ_2	такие, что $P(\theta_1 < p < \theta_2) = \gamma$

*) 1. Вычислить $\alpha = (1+\gamma)/2$

2. Найти по таблице число u_α

3. Вычислить $\varepsilon = u_\alpha \times \mathrm{sqrt}(p^* (1-p^*) / n)$

4. Положить $\theta_1 = p^* - \varepsilon$, $\theta_2 = p^* + \varepsilon$

Задача нахождения доверительного интервала

для вероятности события при большом объеме выборки

послесловие

Квантили нормального распределения N(0,1)

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$\Phi(t) = 1 - \Phi(-t)$

$\Phi(u_\alpha) = \alpha$

?	u_α	?	u_α	?	u_α
.50	0	.91	1.341	.995	2.576
.55	.126	.92	1.405	.999	3.090
.60	.253	.93	1.476	.9995	3.291
.65	.385	.94	1.555	.9999	3.719
.70	.524	.95	1.645	.99995	3.891
.75	.674	.96	1.751	.99999	4.265
.80	.842	.97	1.881	.999995	4.417
.85	1.036	.98	2.054	.999999	4.753
.90	1.282	.99	2.326	.9999999	5.199

$n = 1600,\ p^* = 0.2,\ \gamma = 0.95$

$0.18 < p < 0.22$ с надежностью 0.95

Задача нахождения доверительного интервала

для математического ожидания при большом объеме выборки

Дано	
числа n, γ	n – объем выборки, $n > 30$;
выборка $X_1, \ldots, X_n$	γ – доверительная вероятность;
	EX – математическое ожидание СВ **X;** (обозн.)

Алгоритм*)

числа θ_1 и θ_2	такие, что $P(\theta_1 < EX < \theta_2) = \gamma$

*) 1. Вычислить $\alpha = (1+\gamma)/2$ и найти по таблице число u_α

2. Вычислить выборочные характеристики $\overline{x}$ и S;

3. Вычислить $\varepsilon = S \times u_\alpha / \mathrm{sqrt}(n)$

4. Положить $\theta_1 = \overline{x} - \varepsilon$, $\theta_2 = \overline{x} + \varepsilon$

Задача нахождения доверительного интервала

для среднеквадратичного отклонения

при большом объеме выборки

Дано	
числа n, γ	n – объем выборки;
выборка	γ – доверительная вероятность;
$X_1, \ldots, X_n$	DX – среднеквадратичное отклонение СВ **X;**

Алгоритм*)

числа θ_1 и θ_2	такие, что $P(\theta_1 < DX < \theta_2) = \gamma$

*) 1. Вычислить $\alpha = (1+\gamma)/2$ и найти по таблице число u_α

2. Вычислить выборочные характеристики m_4 и S;

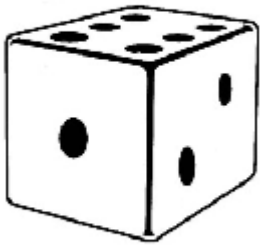
3. Вычислить $\varepsilon = u_\alpha \times \mathrm{sqrt}(m_4 / S^4 - 1) / \mathrm{sqrt}(4 \times n)$

4. Положить $\theta_1 = S \times (1 - \varepsilon)$, $\theta_2 = S \times (1 + \varepsilon)$

Интервальные статистические оценки

пример ($\gamma = 0.95$)

Объем	500	32000
Частоты	0.11 ±0.027	0.10 ±0.003
	0.09 ±0.026	0.10 ±0.003
	0.16 ±0.032	0.15 ±0.004
	0.12 ±0.028	0.15 ±0.004
	0.34 ±0.042	0.30 ±0.005
	0.18 ±0.038	0.20 ±0.004
Выб.средн	5.46 ±0.30	5.46 ±0.037
Выб.скоткл	3.40 ±0.11	3.36 ±0.013



X :	0	0.1
	1	0.1
	2	0.15
	6	0.15
	8	0.3
	9	0.2

EX = 5.5

DX = 11.25

σ = 3.35

Проверка статистических гипотез

Статистическая гипотеза – предположение о свойствах генерального распределения, проверяемые статистическими методами.

Критерий значимости – правило проверки статистической гипотезы.

Статистика критерия – функция **Z : Выборка → Число**, по значениям которой судят о справедливости статистической гипотезы.

Уровень значимости α:

событие с вер. α ≡ событие практически невозможное,
событие с вер. 1-α ≡ событие практически достоверное.

H_0 – проверяемая гипотеза **vs.** H_1 – альтернативная гипотеза.

Критическая область критерия V_k : $\mathbf{P(Z \in V_k \mid H_0) = \alpha}$

Область допустимых значений = **V \ Vk**

≈Схема проверки статистических гипотез

Этап 1. Сформулировать гипотезы H_0 и H_1.

Этап 2. Выбрать уровень значимости α.

Этап 3. Выбрать статистику $Z = Z(\text{Выборка})$ для проверки H_0.

Этап 4. Найти распределение $F(z|H_0)$: хи-квадрат, Стьюдента и др.

Этап 5. Построить критическую область V_k: $P(H_1|H_0) \oplus P(H_0|H_1) \le \alpha$

Этап 6. Вычислить значение $Z_B = Z(\text{Выборка})$.

Этап 7. На уровне доверия $1-\alpha$:
если $Z_B \in V_k$, то H_0 отклоняется.
если $Z_B \in V \setminus V_k$, то H_0 принимается.

Вычисление статистики Z

$\chi^2(n)$ Распределение хи-квадрат с n степенями свободы

– распределение СВ **$X_1*X1 + \ldots + X_n*X_n$**, где $X_i \sim N(0;1)$.

S(n) Распределение Стьюдента с n степенями свободы

– распределение СВ **$X / \mathrm{sqrt}(Y / n)$**, где $X \sim N(0;1)$, $Y \sim \chi^2(n)$

F(n;m) Распределение Фишера с n и m степенями свободы

– распределение СВ **$(X / n) / (Y / m)$**, где $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$

Таблица $\chi^2(n)$ | Таблица (n) | Таблица F(n;m)

Проверка гипотезы о равенстве дисперсий нормально распределенных ген.совокупностей

Выборки	Обе выборки получены независимо
$X_1, \ldots, X_n$	из нормальных ген. совокупностей **X** или **Y** ;
$Y_1, \ldots, Y_m$	$H_0: DX = DY$; $H_1: DX \neq DY$;
число α	α -- уровень значимости

Критерий*)

либо H_0 принимается, либо H_0 отклоняется	

*) 1. Вычислить выборочные несмещаемые дисп. S_X^2 и S_Y^2

2. Вычислить статистику $F_B = S_X^2 / S_Y^2$;

3. Найти по таблице значение $K = F_{1-\alpha/2}(m-1;n-1)$;

4. Если $F_B < K$, то H_0 принимается, иначе H_0 отклоняется

Проверка гипотезы о равенстве мат.ожиданий нормально распределенных ген.совокупностей

Выборки	Обе выборки получены независимо
$X_1, \dots, X_n$ $Y_1, \dots, Y_m$ число α	из нормальных ген. совокупностей **X** или **Y** ; Известно, что **DX = DY** ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; H_0 : **EX = EY** ; ; ; ; ; H_1 : **EX ≠ EY** ; ; ; ; ; ; ; ; ; ; α -- уровень значимости

Критерий*)

либо H_0 принимается, либо H_0 отклоняется	

*) 1. Вычислить выборочные характеристики: $\overline{X}, \overline{Y}, S_X^2, S_Y^2$

2. Вычислить статистику $T_B = \dfrac{\overline{X} - \overline{Y}}{\sqrt{mS_X^2 + nS_Y^2}} \sqrt{\dfrac{mn(m+n-2)}{m+n}}$

3. Найти по таблице значение $K = S_{1-\alpha/2}(m+n-2)$;

4. Если $|T_B| < K$, то H_0 принимается, иначе H_0 отклоняется

Проверка гипотезы о равенстве вероятностей двух событий при больших объемах выборок

Отн.частоты p_A^*, p_B^* число α	p_A^*, p_B^* – относительные частоты событий **A** и **B** ;; получены из выборок большого объема. p_A, p_B – вероятности событий **A** и **B** ; ; ; $H_0 : p_A = p_B$; ; ; ; ; $H_1 : p_A \neq p_A$; ; ; ; ; ; ; ; ; α -- уровень значимости

Упрощенный критерий*) .

Используется доверительный интервал.

либо H_0 принимается, либо H_0 отклоняется	

*) 1. Вычислить доверительный интервал – числа θ_1 и θ_2 :

$$P(\theta_1 < p_A < \theta_2) = 1 - \alpha$$

2. Если $\theta_1 < p_B^* < \theta_2$,

то H_0 принимается, иначе H_0 отклоняется

Проверка гипотезы о законе распределения генеральной совокупности

Выборка $X_1, \ldots, X_n$	$F(x; \theta_1, \ldots, \theta_k)$ – закон распределения ;
Числа $\theta_1, \ldots, \theta_k$	H_0 : распределение ген. совокупности X есть $F(x; \theta_1, \ldots, \theta_k)$; ; ; ; ; ; ; ; ;
Число α	α – уровень значимости

Критерий хи-квадрат Пирсона*) (идея)

либо H_0 принимается, либо H_0 отклоняется	

*) 1. Выборка $\Rightarrow$ Последовательность интервалов $\Delta_1, \ldots, \Delta_m$

2. Выч. n_i – к-во X_j в Δ_i и $p_i = P(x \in \Delta_i)$

3. Вычислить статистику

$$Z = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

4. Если $0 < Z < \chi^2_{1-\alpha}(m-k)$,

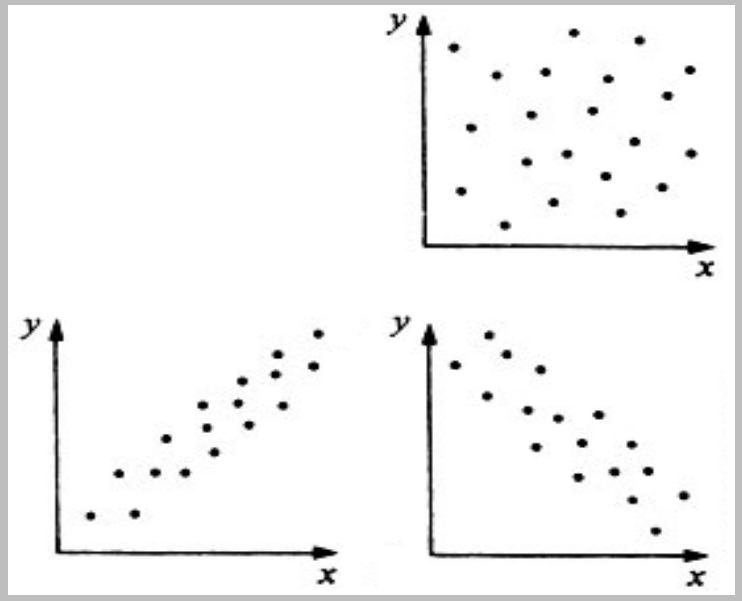
то H_0 принимается, иначе отклоняется

Корреляционный анализ

-- раздел математической статистики, исследующий свойства статистических зависимостей между СВ.

Двумерная выборка

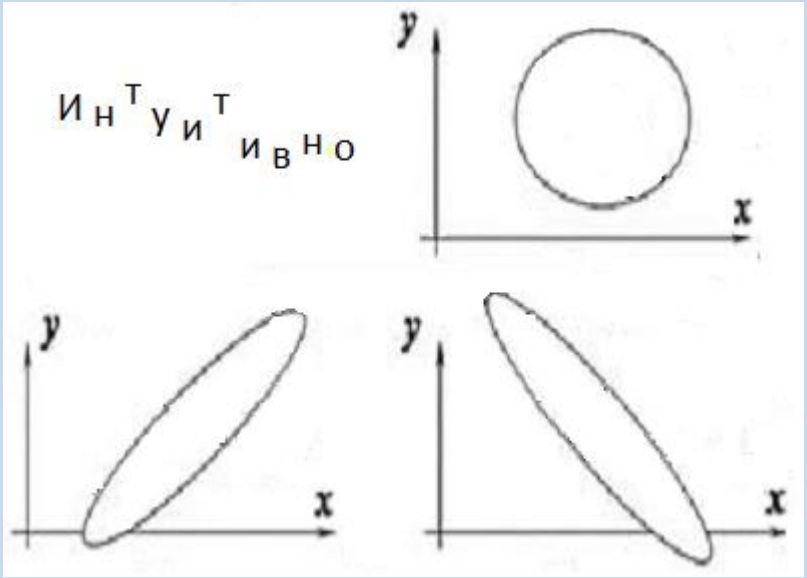
$(X_1,Y_1), \ldots, (X_n,Y_n)$



Генеральная совокупность =

двумерная СВ: $\{ (x_i,y_i) \}$

Распределение: $\{ ((x_i,y_i),p_i) \}$



Порождает СВ X и Y, и вопрос о $\exists$ зависимости между X и Y.

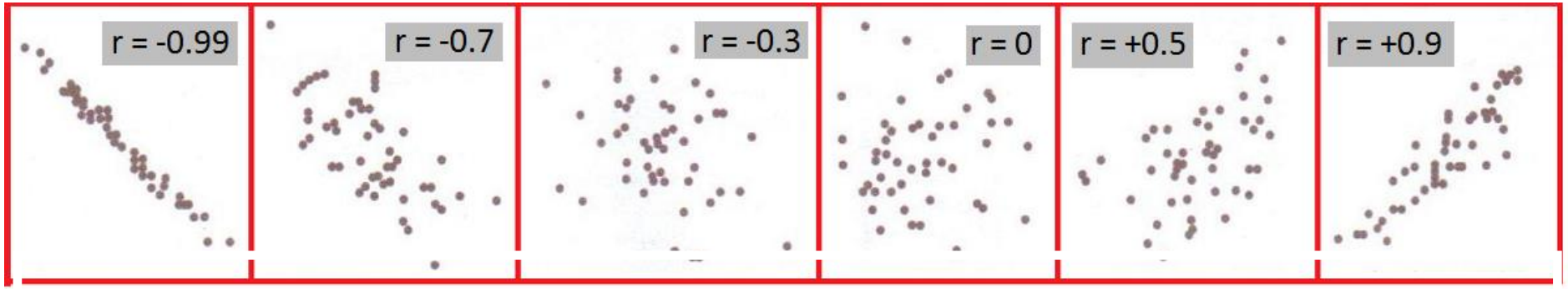
Коэффициент корреляции

$(X_1,Y_1), \ldots, (X_n,Y_n) \Rightarrow$ *Выборочный коэффициент корреляции:*

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

Коэффициент корреляции двух СВ:

$$\rho(X,Y) = \frac{E(X-EX)(Y-EY)}{\sqrt{DX\,DY}}$$



r_{XY} vs. $\rho(X,Y)$

Проверка гипотезы о независимости случайных величин

Выборка	Генеральная совокупность имеет нормальное распределение ;
(X_i,Y_i) i=1..n число α	$H_0: \rho(X,Y) = 0$; ; ; ; $H_1: \rho(X,Y) \neq 0$; ; ; ; ; α -- уровень значимости

Критерий*)

либо H_0 принимается, либо H_0 отклоняется	

*) 1. Вычислить $r = r_{XY}$

2. Вычислить статистику $T_B = \text{sqrt}(n-2)\times r/\ \text{sqrt}(1- r^2)$

3. Найти по таблице значение $K = S_{1-\alpha/2}(n-2)$;

4. Если $|T_B| \leq K$, то **H_0** принимается, иначе **H_0** отклоняется

Грубая проверка гипотезы о независимости случайных величин

Если $|r_{XY}| \times \mathrm{sqrt}(n-1) > 2.5$,

то X, Y зависимы с уровнем значимости 0.005

Если $|r_{XY}| \times \mathrm{sqrt}(n-1) > 3.0$,

то X, Y зависимы с уровнем значимости 0.001

Регрессионный анализ

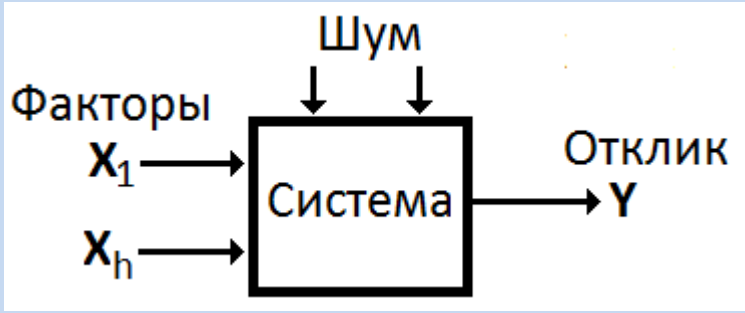
-- раздел математической статистики, исследующий статистические зависимости между двумя и более СВ.

Двумерная выборка

$(X_1,Y_1), \ldots, (X_n,Y_n)$

$y(x) = B_0 + B_1 x$

→ эмпирическая простая линейная регрессия



Зависимость Y от $X_1,\ldots X_h$?

1. $h = 1$
2. Зависимость EY от X, т.е. цель -- поиск функции

   $y(x) = M_x Y$
3. $y(x) = \beta_0 + \beta_1 x$

→ теоретическая простая линейная регрессия

Построение эмпирической простой линейной регрессии

Метод наименьших квадратов:

Дано $(X_1,Y_1), \ldots, (X_n,Y_n)$. Требуются B_1 и B_2 такие, что

$$\Sigma_{i=1..n}(Y_i - B_0 - B_1 x_i)^2 \rightarrow \min$$

Решение: $B_1 = r_{XY}\frac{S_Y}{S_X}$, $B_0 = \overline{Y} - \overline{X}\,B_1$

\+ **Условия Гаусса – Маркова** (1777-1855; 1856-1922)

Отклонения $Y_i - B_0 - B_1 X_i$

(а) нормально распределены,

(б) независимы,

(в) имеют нулевые м.ожидания и равные дисперсии.

Задача нахождения доверительного интервала для коэффициентов простой линейной регрессии

Выборка

$(X_1,Y_1), \ldots,(X_n,Y_n)$;;;

число γ

Выполняются условия Гаусса – Маркова ;;

γ -- доверительная вероятность ;;;;;;;;;;;;;;;

Обозначения коэффициентов регр.:

B_0, B_1 – эмпирические,

β_0, β_1 – теоретические.

Алгоритм $\exists$

числа

δ_0, δ_1

функция

$\delta = \delta(x)$

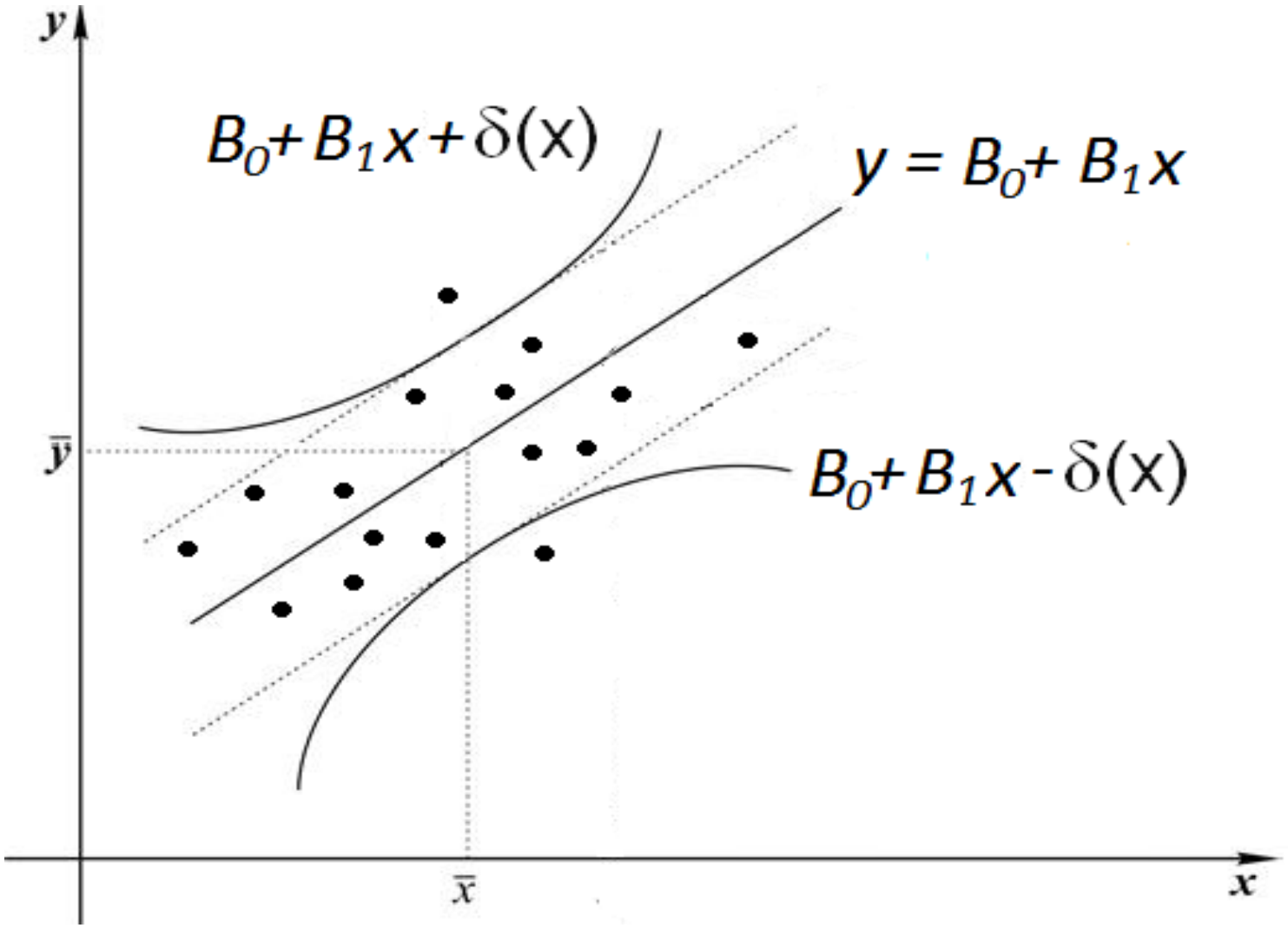
такие, что $P(B_0 - \delta_0 < \beta_0 < B_0 + \delta_0) = \gamma$

$P(B_1 - \delta_1 < \beta_1 < B_1 + \delta_1) = \gamma$

$P(B_0 + B_1 x - \delta < \beta_0 + \beta_1 x < B_0 + B_1 x + \delta) = \gamma$

$\Rightarrow \delta = \delta(x) \Rightarrow$ Коридор ошибок $\Rightarrow$

Коридор ошибок



В о п р о с ы?

soloviev@glossary.ru
